A. M. Heinecke, H. Paul (Hrsg.): Mensch & Computer 2006: Mensch und Computer im Struktur*Wandel*.
München, Oldenbourg Verlag, 2006, S. 19-27

User Assistance and Video-Based Acquisition of Human Action

Karl-Friedrich Kraiss

Institute for Man-Machine Interaction, RWTH Aachen

Abstract

Developments in software and hardware technologies as, e.g. in microelectronics, mechatronics, speech technology, computer linguistics, computer vision, and artificial intelligence are continuously driving new embedded applications for work, leisure, and mobility. Interfaces to such smart systems exploit the same technologies as the said systems themselves. Actual examples for advanced interface design concern user assistance and video-based acquisition of human action for interacting with machines. This paper describes both concepts and presents some recent implementations.

1 Introduction

Human to human conversation works surprisingly hassle-free. Even breakdowns in conversation due to incomplete grammar or missing information are often repaired intuitively. Reasons for this robustness of discourse are mutually available common sense and knowledge about the subject under discussion. Also, the conduct of a conversation follows agreed conventions, allows mixed initiative, and provides feedback of mutual understanding. Focused questions can be asked and answers be given in both directions.

In conventional man-machine interaction little of this kind is yet known. Neither user characteristics nor the circumstances of operation are made explicit; the knowledge about context resides with the user alone. However, if context of use was made explicit, assistance could be provided to the user, similar to that offered to the executive by his personal assistant. In consequence an assisted system is expected to appear simpler to the user than it actually is and will be easier to learn. Handling is made more efficient, safer, or even more pleasant.

Another essential ingredient of human communication is multimodality. We gesture and mimic while talking, even at the phone, when the addressee can not see it. We nod or shake the head or change head pose to indicate agreement or disagreement. We also signal attentiveness by suitable body language, e.g. by turning towards a dialog partner. In so doing conversation becomes comfortable, intuitive, and robust.

It is exactly because of this lack of user assistance and multimodality why current interfaces often fail [1]. This paper therefore gives a short introduction in the concept and implementation of user assistance which relies on techniques such as artificial intelligence, and machine learning. It also addresses a particular aspect of multimodality, i.e. the use of gestures and facial expressions in interfacing. Here computer vision algorithms are needed, which only recently have achieved the maturity for out-of-the-laboratory application. Finally a spectrum of actual applications is presented.

2 The Concept of User Assistance

The generation of assistive functions for interaction support relies on the system architecture presented in (Figure 1). As may be seen, the conventional man machine interaction scheme is augmented by a grey block labelled "user assistance", which takes "context of use" as an input and provides "information display augmentation", "user input augmentation", and "automation" as three kinds of assistive output [2, 3].

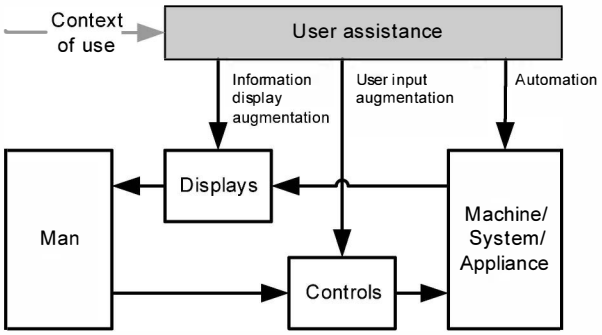

Figure 1: Interaction augmented by user assistance

Context of use identification

Correct identification of context of use is a prerequisite for assistance to be useful. It describes the circumstances under which a system operates. This includes the state of user, the system state, and the situation.

User state refers to the user's identity, whereabouts, workload, skills, preferences, and intentions. Whereabouts are easily accessed by telecommunication and ambient intelligence methods. Workload assessment is based either on user prompting or on physiological data like heartbeat or skin resistance. *System state* characterization is application dependent. In dialog systems the state of the user interface and of single applications and functions are relevant for an assessment. In contrast, dynamics systems are described by state variables, state variables limitations, and resources. Parameters relevant for *situation assessment* are

system dependent and different for mobile or stationary appliances. In general, however, the acquisition of environmental conditions poses no serious problem, as sensors are available abundantly.

As an example consider the context of car driving, where driver state, vehicle state, and traffic situation may be described by the parameters given in (Figure 2).

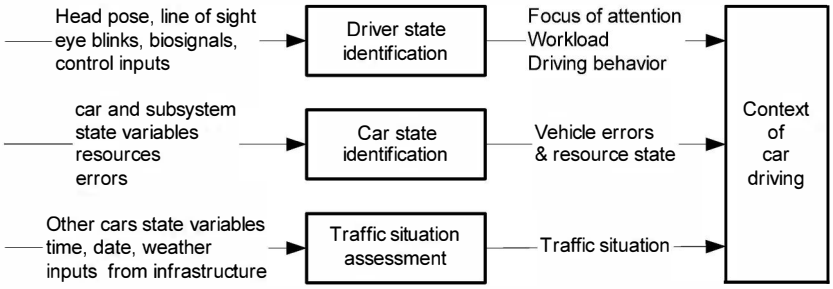

Figure 2: Factors relevant in the context of car driving

Provision of user assistance

As soon as context of use has been identified, assistance can be provided in three modes (see Figure 1). First information display may be improved, e.g. by providing information just in time and coded in the most suitable modality. In case of driving this may relate to improved headlights, predictive information about impending danger, or to the provision of commands and alerts.

Secondly steering and control inputs may be adapted, modified, or limited as the situation requires. Car driving related examples are the anti block brake system (where the force executes at the pedal pressure is limited), or the braking assistance (where the pedal pressure is amplified).

Thirdly manual control may be substituted by automation if the required speed or accuracy of manual inputs is beyond human capabilities. The electronic stabilization program (ESP) in cars makes, e.g. use of combined single wheel braking and active steering to stabilize car yaw angle. In general the driver is not even aware of the fact, that he has been assisted in stabilizing his car.

3 Video-based Acquisition of Human Action

Modalities applicable to interfaces are speech, mimics, gesture, and haptics, which serve information display, user input, or both purposes. Speech recognition has been around for almost fifty years and is available for practical use.

Interest in gesture and mimics is more recent. In fact the first related papers appeared only in the nineties. Early efforts to record mimics and gesture in real time in laboratory setups and in movie studios involved intrusive methods with calibrated markers and multiple cameras. Only recently video-based recognition has achieved an acceptable performance level in out-of-the-laboratory settings. Gesture, mimics, head pose, line of sight and body posture can now be recognized based on video recordings in real time, even under adverse real world conditions. Emotions derived from a fusion of speech, gestures and mimics open the door for yet scarcely exploited emotional interaction. In the following some technological aspects of the video-based acquisition of human actions will be discussed.

3.1 Acquisition of Hand Gestures

Due to the numerous degrees of freedom of the hand, its two dimensional picture is not unique and can not be described by form features alone. Therefore hand localization is mainly based on skin color, which represents a robust and sufficiently invariant feature. In addition the face is taken as a reference point [4].

Figure 3 a.) Input frame; b.) Skin color distribution; c.) Segmented skin colored regions.

In real world settings hand tracking is hampered by skin colored objects in the background, which are mostly static like, e.g. wooden furniture. Since depth information is missing in pictures provided by one camera, such objects can not be distinguished from the user's hands. To compensate this effect, a background model is generated, covering all static objects in a picture frame. A comparison then enables the identification of moving skin colored regions.

The segmented patches in Figure 3 c do not allow a direct identification of the underlying hand posture as quite a number of different options for interpretation exist. This ambiguity is resolved by checking subsequent frames in a picture sequence. Several heuristics are formulated to assign plausibility values to the various available posture hypotheses. Furthermore posture hypotheses are evaluated by making reference to a biometric 2D-sceleton model of the torso and the arms [6]. The validity of various hypotheses is continually logged in parallel until a gesture terminates and all relevant information has been collected. It is only then, that in view of the entire gesture a winning posture is selected.

A further problem results from the fact that during gesturing hands may occlude each other or the face. The almost identical color of hands and face then prevents an effective segmenta-

tion of overlapping regions so that in this case the position of left and right hand can not be identified precisely from one frame. Therefore again a sequence of frames is considered. Tracking is then based on the hand shapes found in the undisturbed views immediately before and after overlap.

The described approach to gesture recognition and hand tracking uses off-the-shelf computer hardware and one Webcam mounted in front of the user. Processing is in almost real time. No markers or data gloves are needed. During extensive testing it proved to work reliably in mobile application, for common backgrounds, and in variable illumination.

3.2 Acquisition of Facial Expressions

For the video-based acquisition of facial expressions the face is first segmented and enlarged by pixel interpolation (Figure 4 a). Noise in the picture as, e.g. shadows on the face and irradiation is then removed by special picture processing. Information about edges, corners, and color distributions is derived in parallel from three false color pictures [4].

Figure 4 a.) Face; b.) Overlaid face graph; c.) 3 D-Head model with texture.

Based on this composed information selected face regions around the eyes and the mouth are localized by matching a face graph iteratively onto an individual face (Figure 4 b). The graph model employs generic knowledge about face texture and face geometry at 70 characteristic points on a face. By localizing these landmarks and based on their relative positions the interesting face regions can then be identified.

Positioning of the face graph may be aggravated by individual difference like a head pose different from frontal, wearing of binoculars and beards, or long hairs covering the eyes. To handle these problems the face model has to be matched to each individual. To this end a virtual biometric 3D-head is calculated on which a frontal facial view is mapped with correct geometry and texture (Figure 4 c.). The simulated head is subsequently used to generate reference faces for varying head poses and illuminations. Since the biometric 3D-head features also an anatomically correct muscle model, different facial expressions can be generated synthetically and stored for adaptive face graph training.

Following the successful positioning of the face graph single facial features like iris and eyebrow position, eye blinks, or mouth contour are determined as exactly as possible by the

combined application of a variety of specialized picture processing algorithms. For the final facial expression analysis the single identified features are synthesized and coded into facial action units.

3.3 Applications of Hand Gesture and Facial Expression Commands

In spite of the fact, that video-based recognition of hand gestures and facial expressions has only recently reached an acceptable performance, a wide spectrum of applications has already evolved; some products even have successfully reached the marketplace.

Substitution of Data Gloves

Interaction with virtual reality mostly involves data gloves of varying technology. Vision based acquisition of hand and finger posture may substitute data gloves in the near future. Since the 2D projection of a hand resulting from one camera is ambiguous and does not allow unique posture identification, the user must wear a cotton glove with colored fingertips (Figure 5). Nevertheless this solution is much less intrusive than common data gloves.

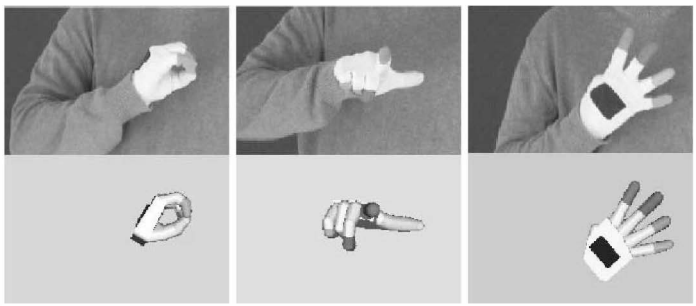

Figure 5: Various hand postures (upper part) and their identification with a hand computer model (lower part). The user wears a cotton glove with colored finger tips [5].

Gesture control of dialog systems in cars

The functionality of dialog systems in cars grows exponentially. To facilitate the handling during driving, multimodal user interfaces have been developed which make use of several sensory and motor channels of humans. Activation of knobs and dials requires allocation of visual attention. Therefore speech is widely used, since it does not load the visual channel. However, in case of environmental noise, speech recognizers fail. Gesture control as depicted in Figure 6 a.) offers a solution to this problem (the camera is mounted behind the rear mirror).

Sign language recognition

Sign language are fully-fledged language fort he daily communication between and with the deaf. The mitigation of linguistic contents is based on manual and non manual means of expression. Automated sign language recognition will improve the communication between the hearing and the deaf population. In Figure 6 b.) a recognition system consisting of a lap-top and one webcam mounted on it is depicted, which is able to recognize about 250 gestures in near real time.

Figure 6: (a) Gesture control of dialog systems in cars [6]; Sign language recognition [7].

Controls for people with severe motor handicaps

For people with severe motor handicaps like paraplegics head motions and facial expression may be the last resort to enable interaction with the environment. Recently a wheelchair has been developed, that is controlled by head pose, eye point of regard, and mouth shape [8]. The face of the wheelchair driver is illuminated by infrared light and recorded by a webcam.

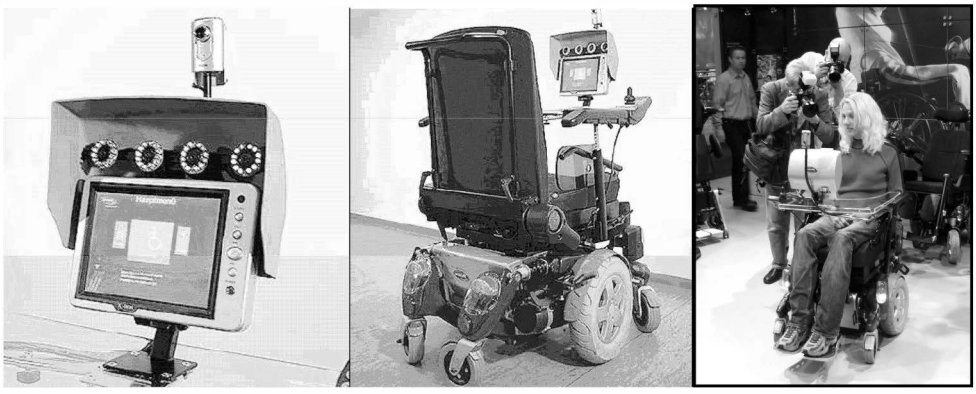

Figure 7: Wheelchair controlled by head pose and facial expressions [8].

Driver status acquisition

Video-based acquisition of the eye-blinks, lip movements and head pose of a driver is part of driver status acquisition which is desirable in various driver assistance systems as, e.g., automatic heading control, speed control, distance keeping, and stop- and-go [8].

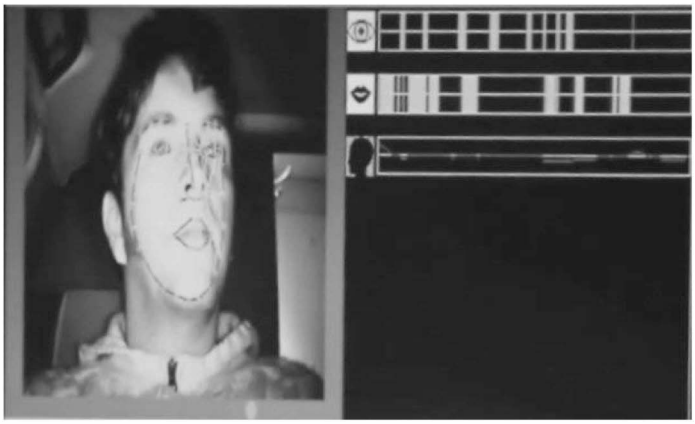

Figure 8: Video-based acquisition of eye blinks, lip movements and head pose for driver status acquisition. Left: the drivers face with overlaid face graph. Right: the logged protocol data.

3.4 Conclusions

In this paper two novel approaches for advanced interface design where discussed. First the concept of user assistance and its implementation were dealt with. Some examples from the realm of car driving assistance were given. Then it was shown how gestures and facial expressions can be acquired with a camera and used for interacting with machines in real world settings. This was illustrated by a variety of recently implemented application examples. From this it appears that user assistance, hand gesture commands and facial expressions commands are essential add-ons to advanced interfaces, which tend to improve usability and reliability of interaction with machines.

References

[1] Kraiss K.-F. (Ed.) (2006): *Advanced Man-Machine Interaction*, Springer Verlag.

[2] Kraiss K.-F. (2006): Assisted Man-Machine Interaction In: Kraiss K.-F. (Ed.) *Advanced Man-Machine Interaction*, Springer Verlag.

[3] Libuda L.; Kraiss K.-F. (2003): Dialogassistenz im Kraftfahrzeug. In: *45. Fachausschusssitzung Anthropotechnik der DGLR „Entscheidungsunterstützung für die Fahrzeug- und Prozessführung*," Volume DGLR-Bericht 2003-04, pp. 255-270, 14.-15. Oktober, Neubiberg.

[4] Zieren, J.; U. Canzler (2006): Non-intrusive Acquisition of Human Action. In: Kraiss K.-F. (Ed.) *Advanced Man-Machine Interaction*, Springer Verlag.

[5] Zieren J.; Dick, T; Kraiss, K.-F. (2006): Visual Hand Posture Recognition in Monocular Image Sequences. DAGM, Springer LNCS, to appear.

[6] Akyol S.; Canzler U.; Bengler K.; Hahn W. (2000): Gesture Control for Use in Automobiles In: *Proc. of the IAPR MVA 2000 Workshop on Machine Vision Applications*, pp. 349-352, November 28-30, Tokyo.

[7] Zieren J.; Kraiss, K.-F. (2004): Non-Intrusive Sign Language Recognition for Human-Computer Interaction. In: *9th IFAC/IFIP/IFORS/IEA Symposium Analysis, Design, and Evaluation of Human-Machine Systems*, pp. CD-paper 49, September 7-9, Atlanta, Georgia

[8] http://www.cancontrols.com/